Recycle Bin
Microsoft Edge
Orange
RapidMiner Studio
PortableApps
RapidMiner
Orange-dat...

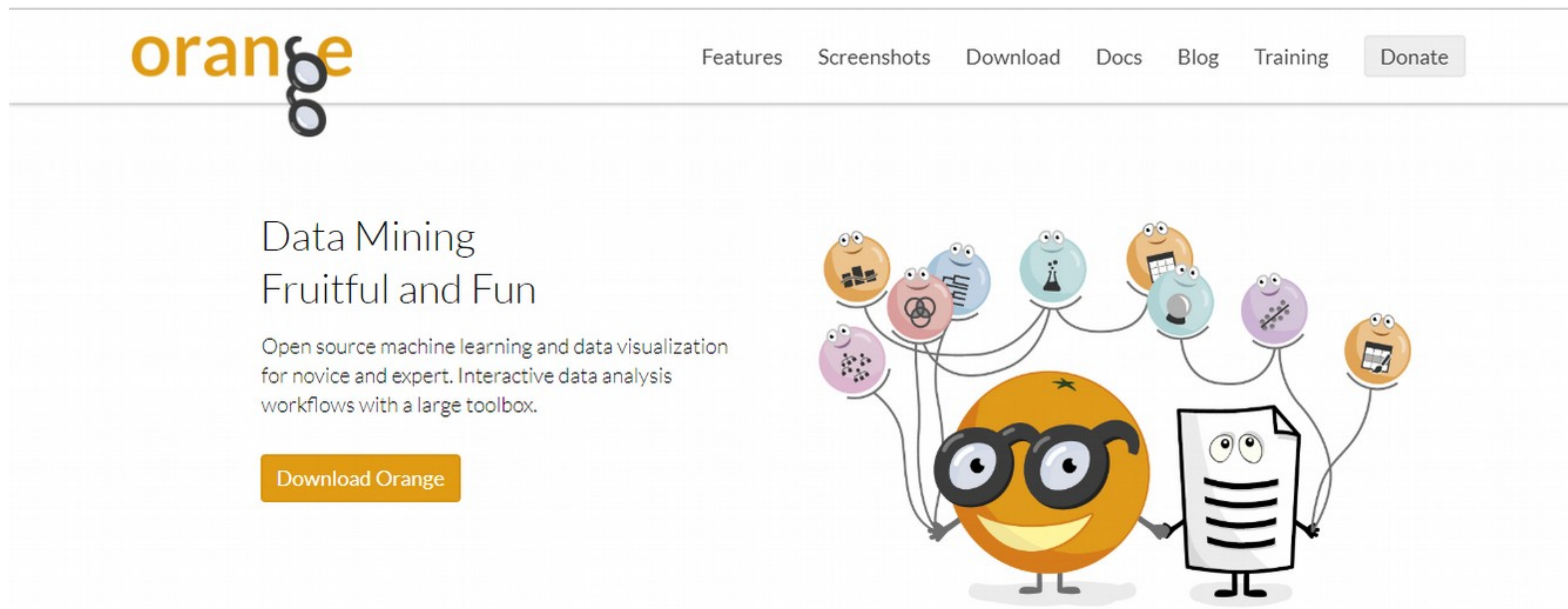
Minería de Datos para economistas y profesionales de negocios

Un viaje en el mundo de los datos sin programación (códigos) con el software Orange. Serán tratados importantes temas como: Técnicas de agrupamiento, modelos predictivos, minería de texto y análisis de imágenes.

Indice del Taller

Introducción e Instalación 3
Image Analytics: Clustering 8
Predecir la elegibilidad de un préstamo 15
Trabajar con Texto 29

Introducción e Instalación



Con la creciente uso de datos, necesitamos herramientas que eliminen la dificultad de realizar operaciones - cálculos con datos. No todos están dispuestos a aprender codificación, aunque quisieran aprender y aplicar la los datos. Aquí es donde una herramienta con un entorno grafico puede ser muy útil.

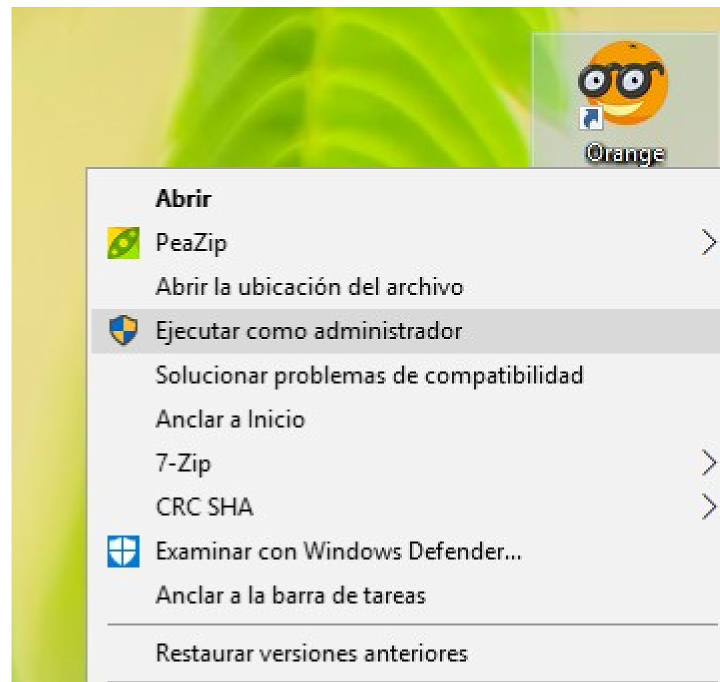
Orange: La herramienta ideal para principiantes que desean visualizar patrones y entender sus datos sin saber de codificar.

Configurando tu sistema

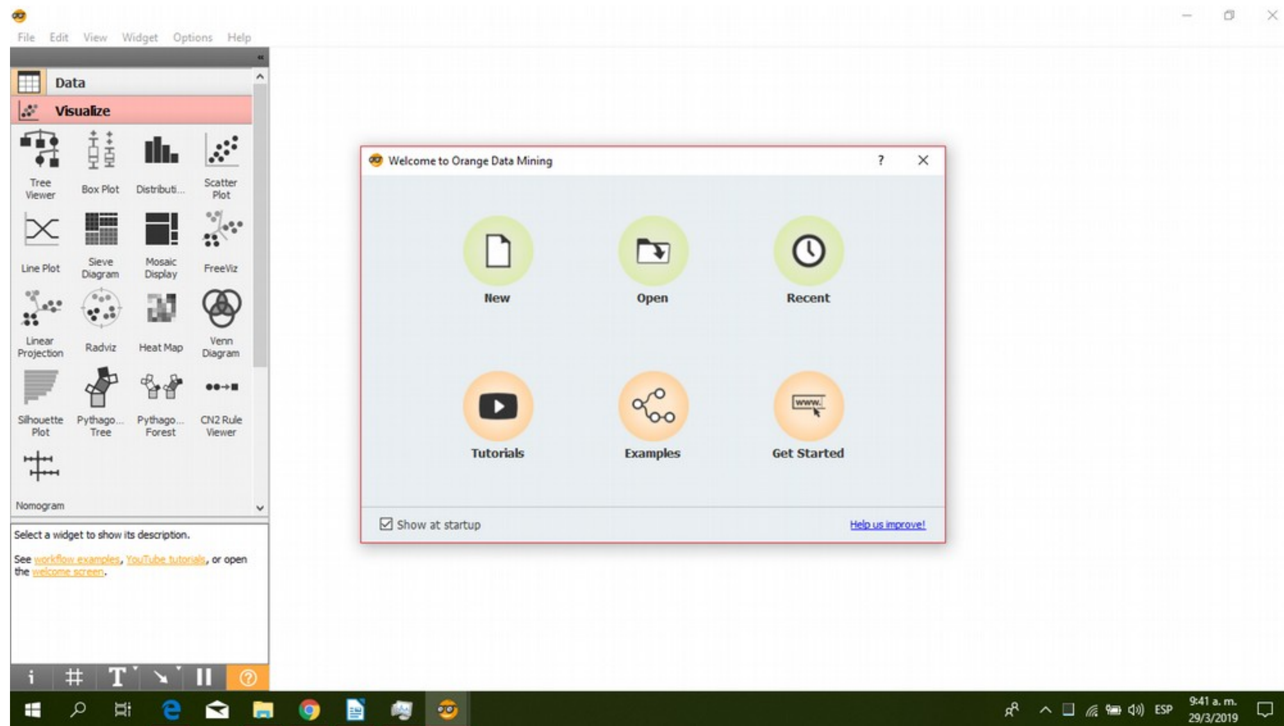
https://orange.biolab.si
y seleciona Download
seleciona tu sistema Windows o Apple OSX

https://drive.google.com/drive/folders/1OLkd4tZMnjuofLTXOY94EwKs9fhLVT0Z?usp=sharing
extención .exe es Windows / dmg es para Apple OSX

Instale la plataforma y configure el directorio de trabajo para que Orange almacene sus archivos.
Ejecutar como administrador su software Orange!

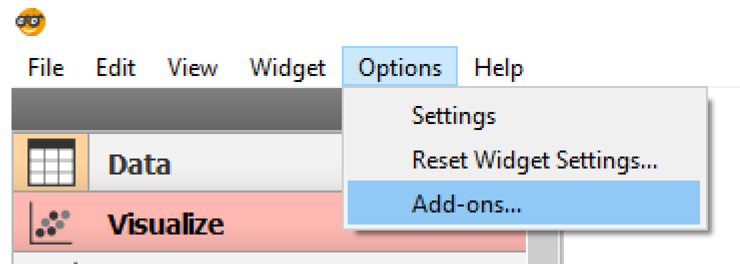


Nos veamos el pantalla de inicio.

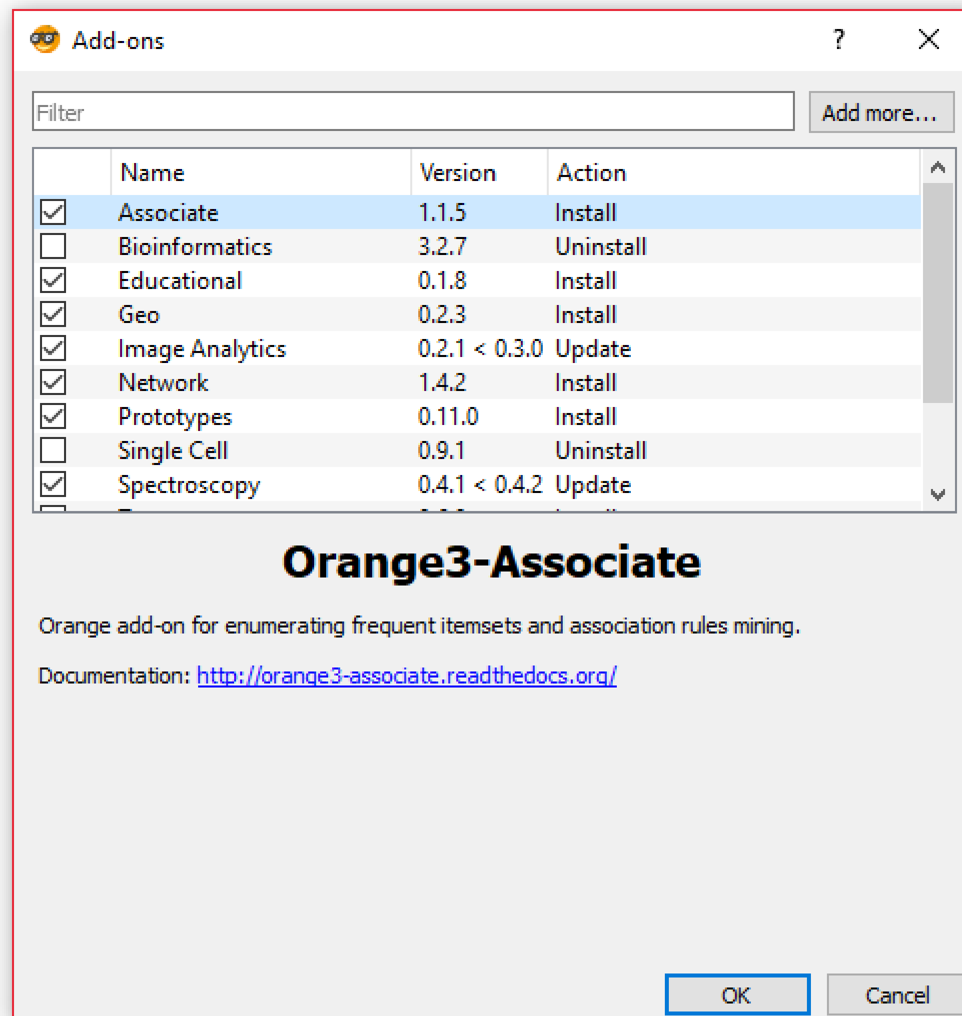


Cerrando el pantalla de bienvenida en el X.

Abrimos Options Add-ons ..



Aquí veamos este menú de selección, selecciona **Image Analytics**, y **Text,** y confirmamos con el botón “OK” la instalación de “Add-ons”.



Después de **reiniciar la aplicación** estamos listo para trabajar.

El espacio de trabajo

Así es como se ve la página de inicio de Orange. Los **objectos se mover con ”Arrastrar y soltar” (drag y drop)** de caja de herramientas a nuestro proyecto.

Un **widget es el punto de procesamiento básico de cualquier manipulación de datos**.

Puede realizar una serie de acciones en función de lo que elija en el selector de widgets a la izquierda de la pantalla.

Un **flujo de trabajo es la secuencia de pasos o acciones es** para realizar una tarea en particular.

Image Analytics: Clustering
Análisis de imágenes: Agrupamiento

Material: **Software Orange,**

Fichero de datos: **Directorio “domestic-animals”, o archivo “domestic-animals.zip”**

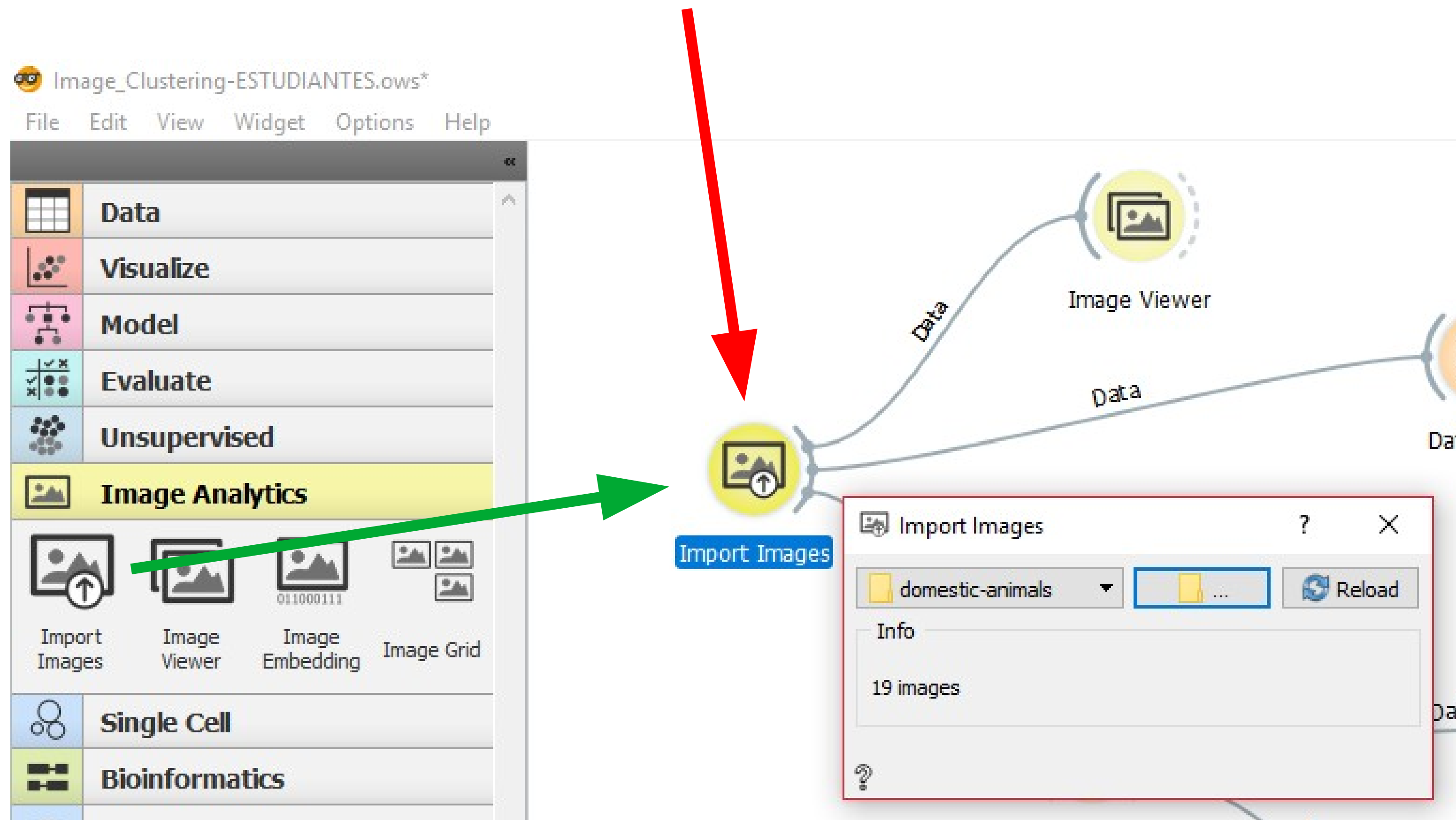
Los datos no siempre vienen en una forma de una tabla. También puede ser una colección de texto, grabaciones de audio, materiales de video o incluso imágenes. Sin embargo, las computadoras solo pueden trabajar con números, por lo que para cualquier extracción de datos, necesitamos transformar esos datos no estructurados en una representación vectorial.

Para recuperar números de datos no estructurados, Orange se puede incrustar en la red. Acabamos de comenzar a incluir varios modolus en Orange, que están disponibles para texto e imágenes.

Aquí, damos un ejemplo de incrustación de imagen y mostramos lo fácil que es usarlo. Técnicamente, Orange enviaría la imagen al servidor, donde el servidor empujaría una imagen a través de una red neuronal profunda pre-entrenada, como el **Inception v3** de Google. La mayoría de las redes profundas se entrenaban con algún propósito especial en mente. Inception v3, por ejemplo, puede clasificar imágenes en cualquiera de 1000 clases de imágenes. Podemos ignorar la clasificación, considerar la penúltima capa de la red con 2048 nodos (números) y usarla para la representación basada en vectores de la imagen.

Aquí tenemos **19 imágenes de animales domésticos (Directorio "domestic-animals", o archivo "domestic-animals.zip")**. Primero, descargue las imágenes y descomprímalas. Luego use el widget Importar imágenes del complemento Image Analytics de Orange y abra el directorio que contiene las imágenes.

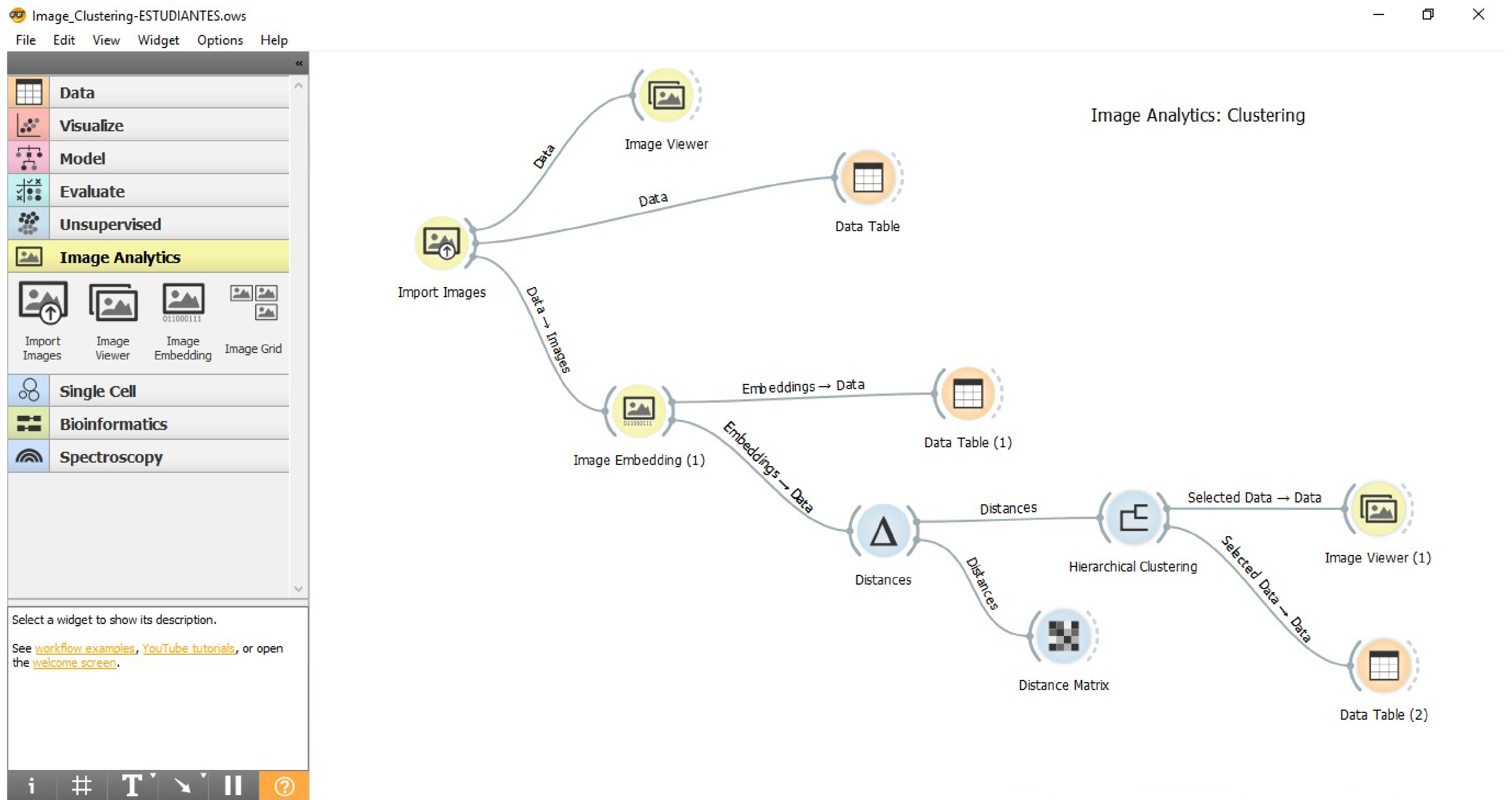
Arrastramos **Import Imagenes** a nuestro espacio de trabajo, y con un "doble click" cargamos nuestro directorio con los imágenes previamente descargado.



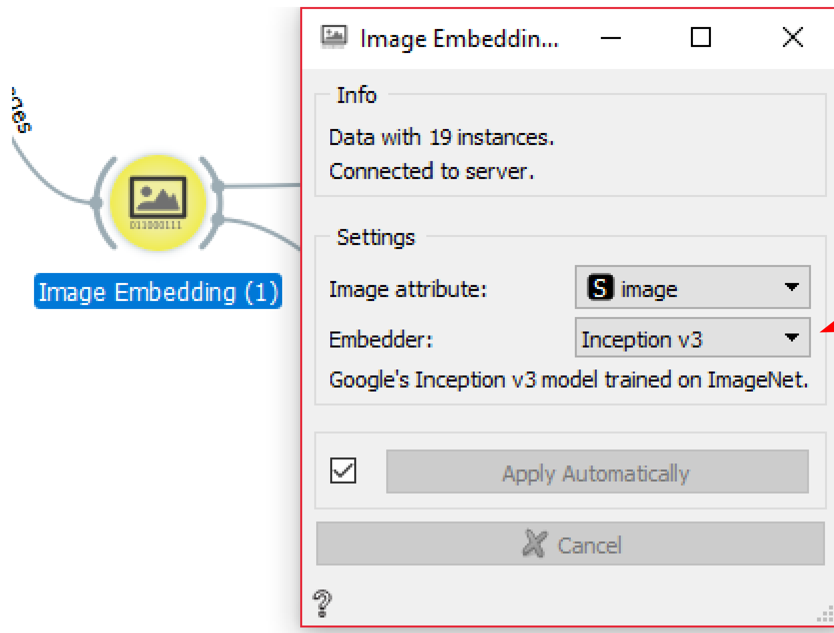
Instalamos los modulos (wigets) **Image Viewer** y **Data Table**. Podemos visualizar imágenes en el widget Visor de imágenes.

Con un doble click en **Image Viewer**, se puede vea los imágenes, con **Data Table** la tabla Completamos nuestra aplicación, como se vea en esta grafica.

Completamos nuestra aplicación como veamos en esta grafica.

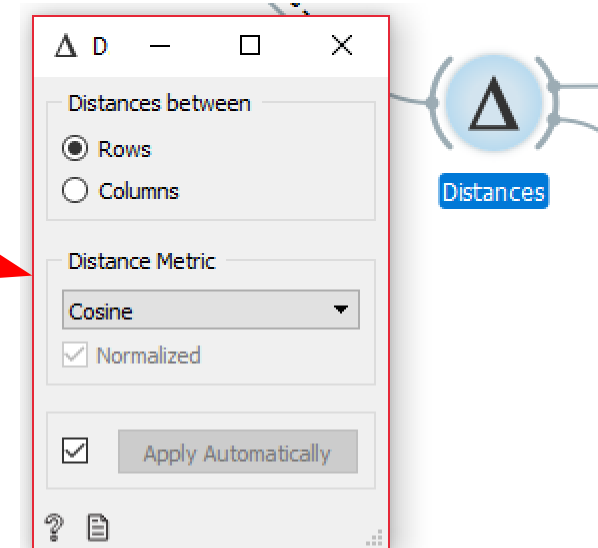


Ahora necesitamos configurar nuestra aplicación.

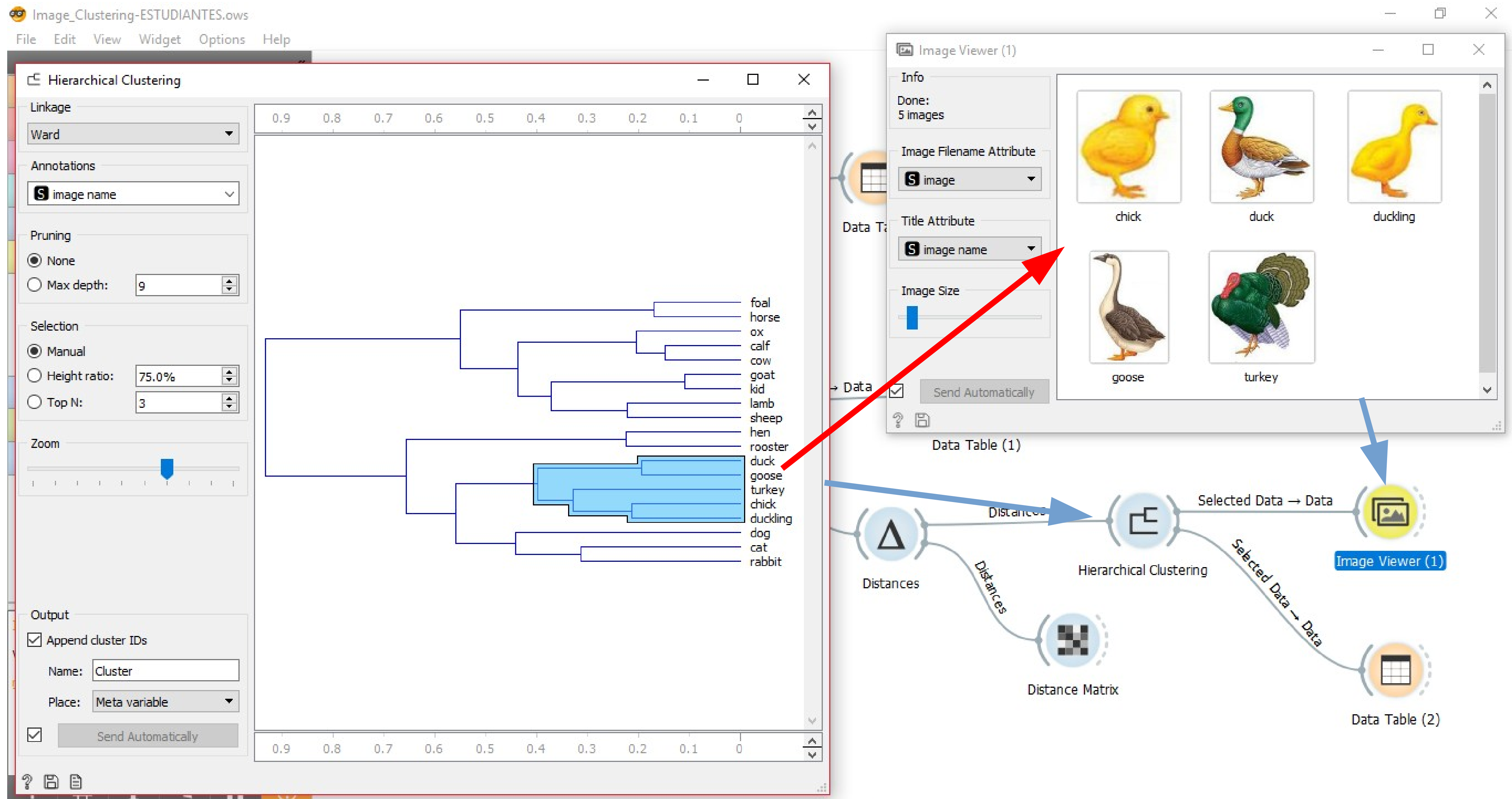


Con un doble click abrimos **Image Embedding** y selecionamos **Inception v3** y cerramos

En el modulo **Distances** necesitamos seleccionar **Cosine**



En **Hierarchial Clustering** veamos el dendrograma y en el **Image Viewer** los objetos selecionados en el dendograma.



¡Imágenes tienen números para calcular!

Predecir la elegibilidad de un préstamo

Material: **Software Orange,**

Fichero de datos: **archivo “Credito-Train-es.csv”**

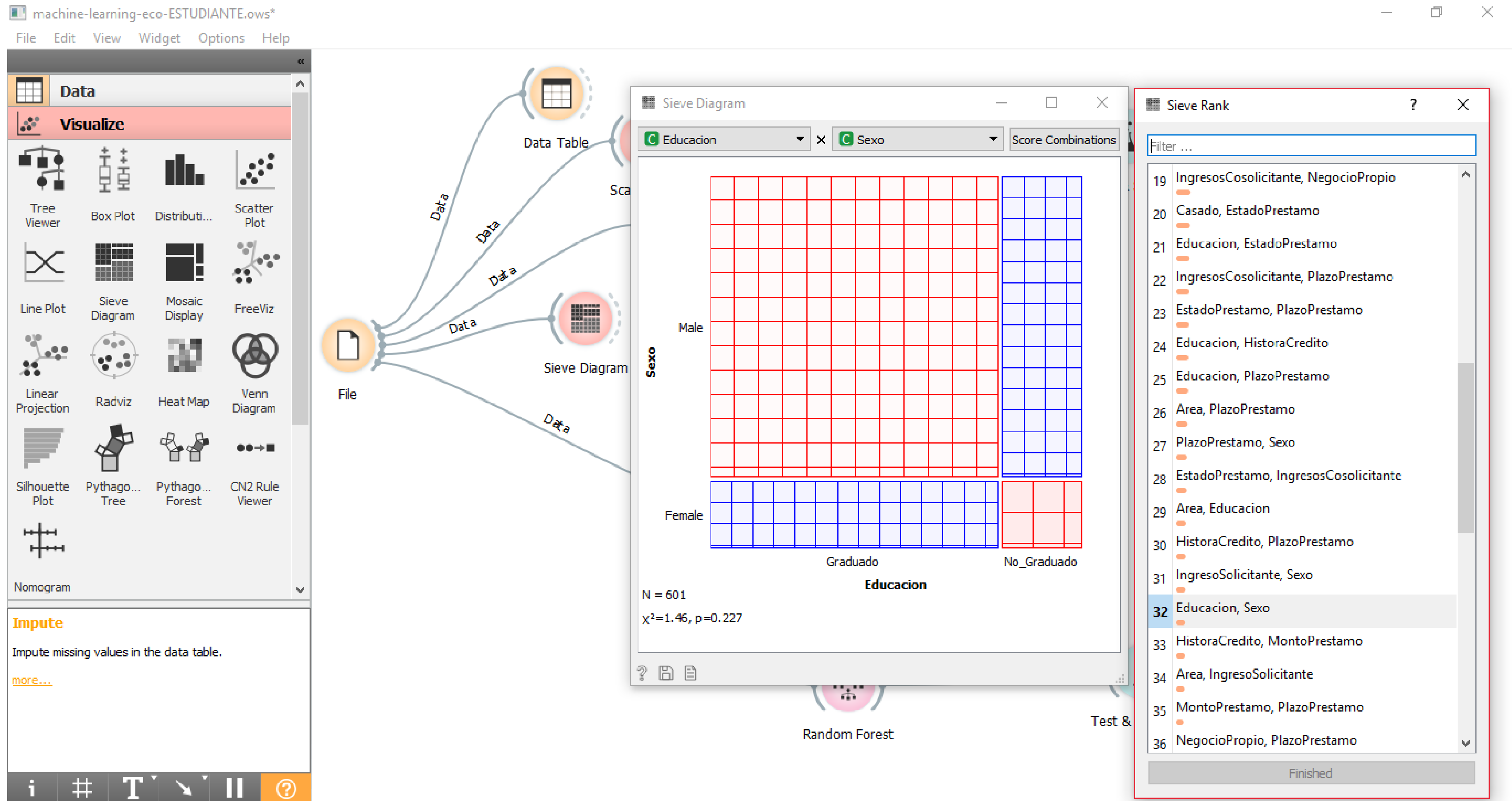
Importando los archivos de datos

Paso 1: haga clic en la pestaña "Datos" en el menú de selección de widgets y arrastre el widget "Archivo" a nuestro flujo de trabajo en blanco.

Paso 2: Haga doble clic en el widget "Archivo" y seleccione el archivo **“Credito Train es.csv”** para cargar en el flujo de trabajo.

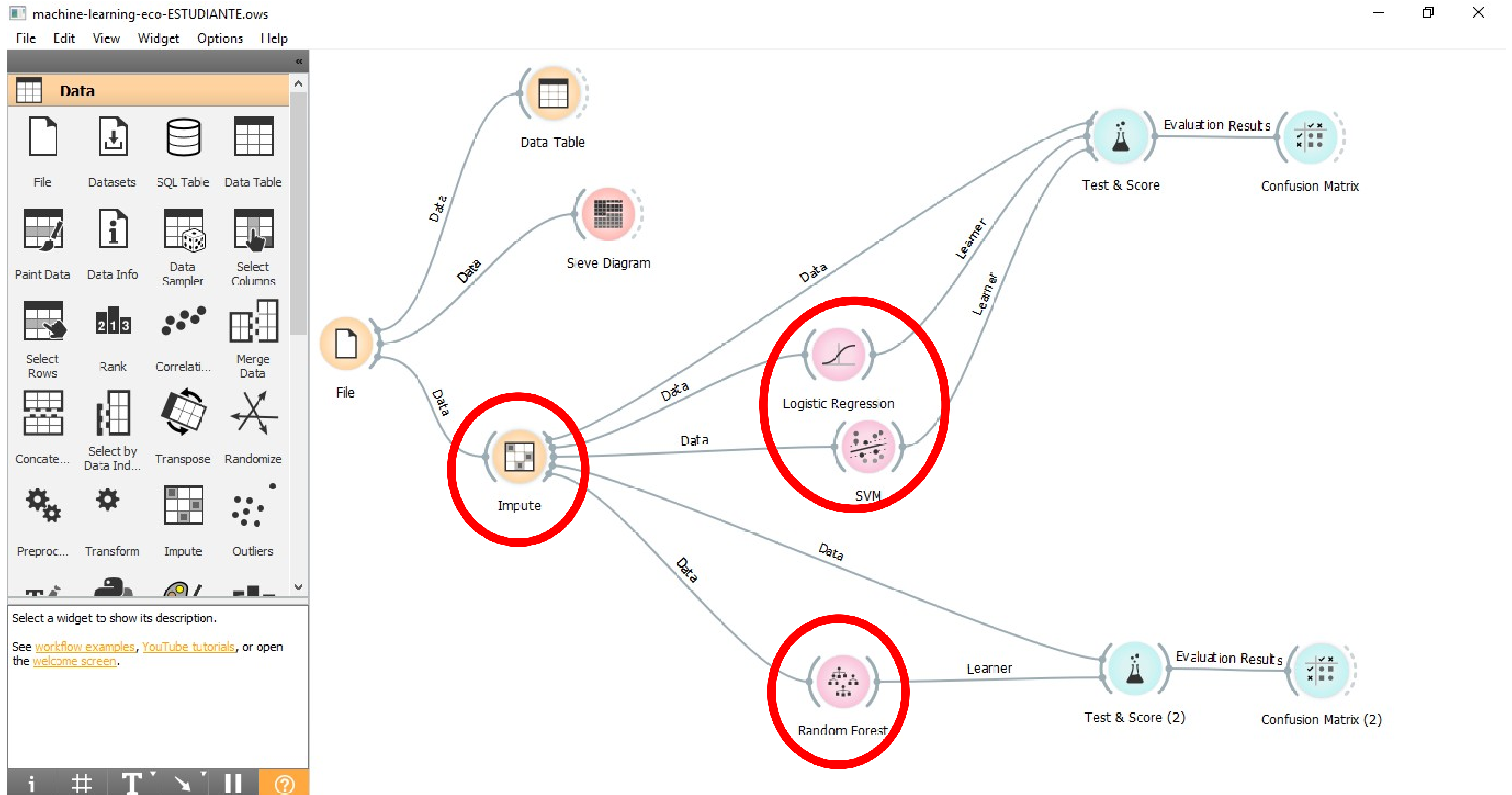
Ahora, haga doble clic en la **columna EstatoPrestamo** y selecciónela como la **variable de destino (target)**. Haga clic en **Aplicar (Apply)**.

Visualizar nuestros Datos



En este ejemplo usamos el el **“Sieve Diagram”.**
Podemos vea también con **“Data Table”** nuestros datos en forma de tabla.

Creamos nuestro modelo



Modulo Impute

Reemplaza valores desconocidos en los datos.

Entradas
- Datos: entrada de datos
- Alumno: algoritmo de aprendizaje para la imputación.

Salidas
- Datos: conjunto de datos con valores imputados.

Algunos algoritmos y visualizaciones de Orange no pueden manejar valores desconocidos en los datos. Este widget hace lo que los estadísticos llaman imputación: sustituye los valores perdidos por valores calculados a partir de los datos o establecidos por el usuario. La imputación predeterminada es (1-NN)

Más detalles en: https://en.wikipedia.org/wiki/Imputation_(statistics)

Logistic Regression / Regresión logística

El algoritmo de clasificación de regresión logística con LASSO (L1) o Ridge (L2) regularización.

Entradas

- Datos: entrada de datos
- Preprocesador: método (s) de preprocesamiento

Salidas

- Aprendiz (Learner): algoritmo de aprendizaje de regresión logística
- Modelo: modelo entrenado
- Coeficientes: coeficientes de regresión logística.

Regresión logística aprende un modelo de Regresión logística a partir de los datos. Solo funciona para tareas de clasificación.

Mas en: https://es.wikipedia.org/wiki/Regresión_logística

Support Vector Machines – SVM / Máquinas de soporte vectorial
Las máquinas de soporte vectorial asignan entradas a espacios de entidades de dimensiones superiores.

Entradas
- Datos: entrada de datos
- Preprocesador: método (s) de preprocesamiento

Salidas
- Aprendiz (Learner): algoritmo de aprendizaje de regresión lineal
- Modelo: modelo entrenado
- Vectores de soporte: instancias utilizadas como vectores de soporte

La máquina de máquinas de soporte vectorial (SVM) es una técnica de aprendizaje automático que separa el espacio de atributos con un hiperplano, maximizando así el margen entre las instancias de diferentes clases o valores de clase. La técnica a menudo produce resultados de rendimiento predictivo supremos. Orange incorpora una implementación popular de SVM del paquete LIBSVM. Este widget es su interfaz gráfica de usuario.

Para las tareas de regresión, SVM realiza una regresión lineal en un espacio de entidad de alta dimensión utilizando una pérdida insensible a ε. Su precisión de estimación depende de una buena configuración de los parámetros C, ε y kernel. El widget genera predicciones de clase basadas en una Regresión SVM.

Más en: https://es.wikipedia.org/wiki/Máquinas_de_vectores_de_soporte

Random forest / Bosque Aleatorio

Predecir utilizando un conjunto de árboles de decisión.

Entradas
- Datos: entrada de datos
- Preprocesador: método (s) de preprocesamiento

Salidas
- Aprendiz: algoritmo de aprendizaje al azar del bosque
- Modelo: modelo entrenado

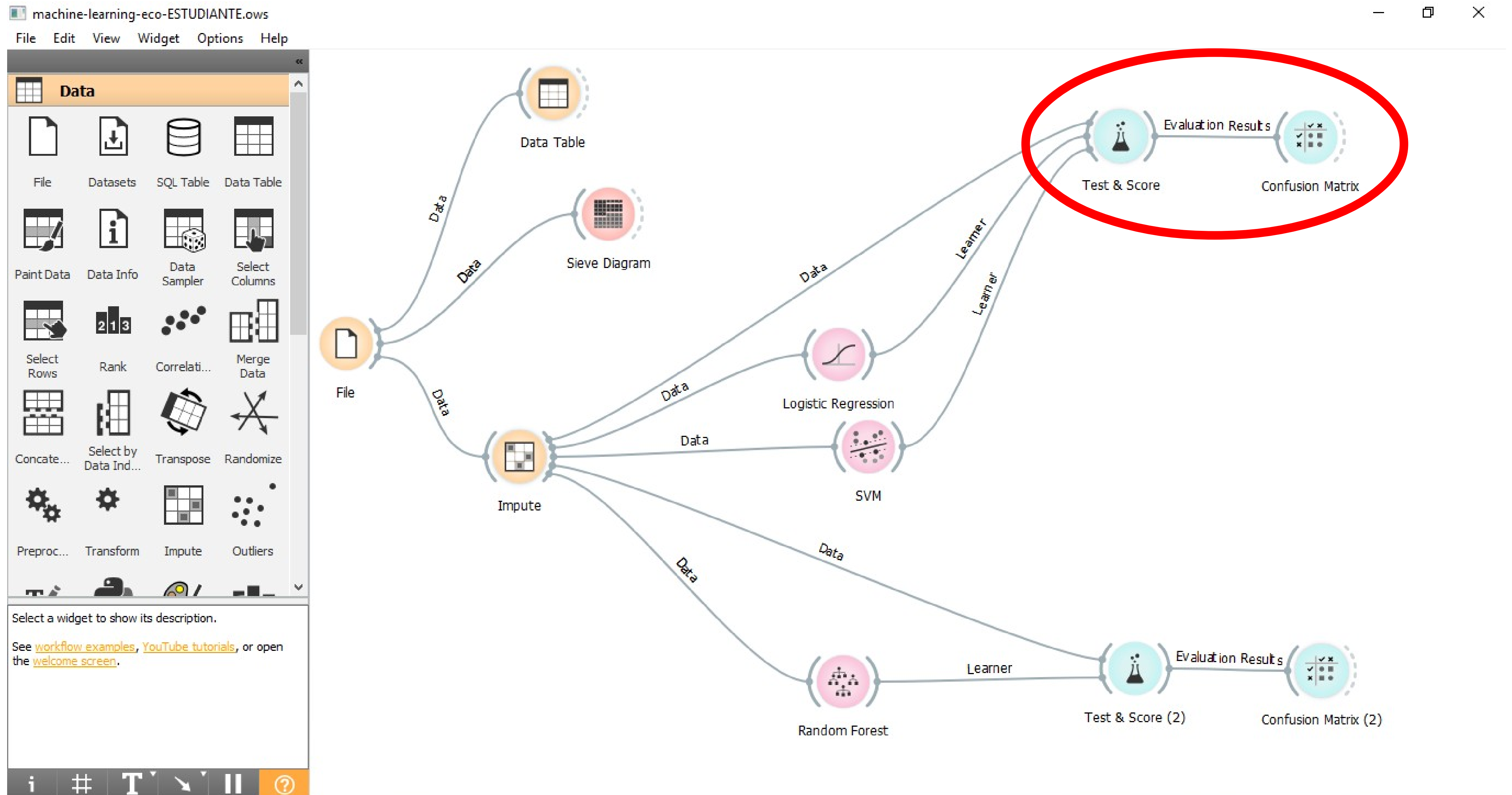
El bosque aleatorio es un método de aprendizaje en conjunto utilizado para la clasificación, regresión y otras tareas. Fue propuesto por primera vez por Tin Kam Ho y posteriormente desarrollado por Leo Breiman (Breiman, 2001) y Adele Cutler.

Random Forest construye un conjunto de árboles de decisión. Cada árbol se desarrolla a partir de una muestra bootstrap a partir de los datos de entrenamiento. Al desarrollar árboles individuales, se dibuja un subconjunto arbitrario de atributos (de ahí el término "Aleatorio"), desde el cual se selecciona el mejor atributo para la división. El modelo final se basa en el voto mayoritario de árboles desarrollados individualmente en el bosque. Se observa en datos con mucho “ruido” no esta preciso.

Random Forest funciona tanto para tareas de clasificación como de regresión.

Más en: https://es.wikipedia.org/wiki/Random_forest

Nuestro Modelo



Test and Score / Prueba y puntuación

Pruebas de algoritmos de aprendizaje en datos.

Entradas

- Datos: entrada de datos
- Datos de prueba: datos separados para la prueba
- Aprendiz: algoritmo (s) de aprendizaje

Salidas

- Resultados de la evaluación: resultados de los algoritmos de clasificación de prueba

El widget prueba los algoritmos de aprendizaje. Se encuentran disponibles diferentes esquemas de muestreo, incluido el uso de datos de prueba separados. El widget hace dos cosas. Primero, muestra una tabla con diferentes medidas de rendimiento del clasificador, como la precisión de la clasificación y el área bajo la curva. En segundo lugar, genera resultados de evaluación, que pueden ser utilizados por otros widgets para analizar el rendimiento de los clasificadores, como el análisis ROC o la matriz de confusión.

La señal del alumno (Learner) tiene una propiedad poco común: se puede conectar a más de un widget para probar a múltiples alumnos (Learner) con los mismos procedimientos.

Test & Score

Sampling

Cross validation

Number of folds: 20

Stratified

Cross validation by feature

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
SVM	0.774	0.808	0.791	0.810	0.808
Logistic Regression	0.771	0.741	0.677	0.784	0.741

AUC: El área bajo ROC* es el área bajo la curva operativa del receptor.

CAL: a precisión de la clasificación es la proporción de ejemplos clasificados correctamente.

F-1: Es un medio armónico ponderado de precisión y recuperación (ver más abajo).

Precission: La precisión es la proporción de verdaderos positivos entre los casos clasificados como positivos, p. Ej. La proporción de Iris virginica correctamente identificada como Iris virginica.

Recall: La recuperación es la proporción de verdaderos positivos entre todas las instancias positivas en los datos, por ejemplo, El número de enfermos entre todos los diagnosticados como enfermos.

* **Receiver operating characteristic curve**, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The ROC curve is created by plotting the true **positive rate against the false positive rate** at various threshold settings. The **true-positive rate** is also known as **sensitivity, recall or probability of detection in machine learning**.

The **false-positive rate** is also known as the **fall-out or probability** of false

machine-learning-eco-ESTUDIANTE.ows

File Edit View Widget Options Help

Data

File, Datasets, SQL Table, Data Table, Paint Data, Data Info, Data Sampler, Select Columns, Select Rows, Rank, Correlati..., Merge Data, Concate..., Select by Data Ind..., Transpose, Randomize, Preproc..., Transform, Impute, Outliers

Confusion Matrix

Display a confusion matrix constructed from the results of classifier evaluations.

more...

Data Table, File, Impute, Test & Score, Evaluation Results, Confusion Matrix

Confusion Matrix

Logistic Regression
SVM

Show: Proportion of predicted

Predicted

Actual	N	S	Σ
N	82.5 %	19.6 %	192
S	17.5 %	80.4 %	422
Σ	114		614

actual: S
predicted: S

Output

Predictions Probabilities

Send Automatically

Select Correct Select Misclassified Clear Selection

Confusion Matrix / Matriz de confusión

En el campo de la inteligencia artificial una matriz de confusión es una herramienta que permite la **visualización del desempeño de un algoritmo** que se emplea en aprendizaje supervisado. Cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real. Uno de los beneficios de las matrices de confusión es que facilitan ver si el sistema está confundiendo dos clases.

Predicted

Actual

	N	S	Σ
N	82.5 %	19.6 %	192
S	17.5 %	80.4 %	422
Σ	114	500	614

	Real negativo	Real positivo
Test negativo	**Correcto Positivo**	**Falso Positivo**
Test positivo	**Falso Negativo**	**Correcto Negativo**

<u>Ejempo: Pacientes Sanos y Enfermos</u>

Correcto positivo: el paciente está enfermo y la prueba lo ha indicado correctamente.
Falso negativo: el paciente está enfermo, pero la prueba lo ha clasificado como sano por error.
Falso positivo: el paciente está sano, pero la prueba lo ha clasificado erróneamente como enfermo.
Correcto Negativo: el paciente está sano y la prueba lo ha indicado correctamente.

Detalles: https://en.wikipedia.org/wiki/Confusion_matrix
https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers

Los Resultados en una tabla

Data Table (1)

Info

402 instances
10 features (no missing values)
Discrete class with 2 values (no missing values)
3 meta attributes (0.7% missing values)

Variables

- [x] Show variable labels (if present)
- [x] Visualize numeric values
- [x] Color by instance classes

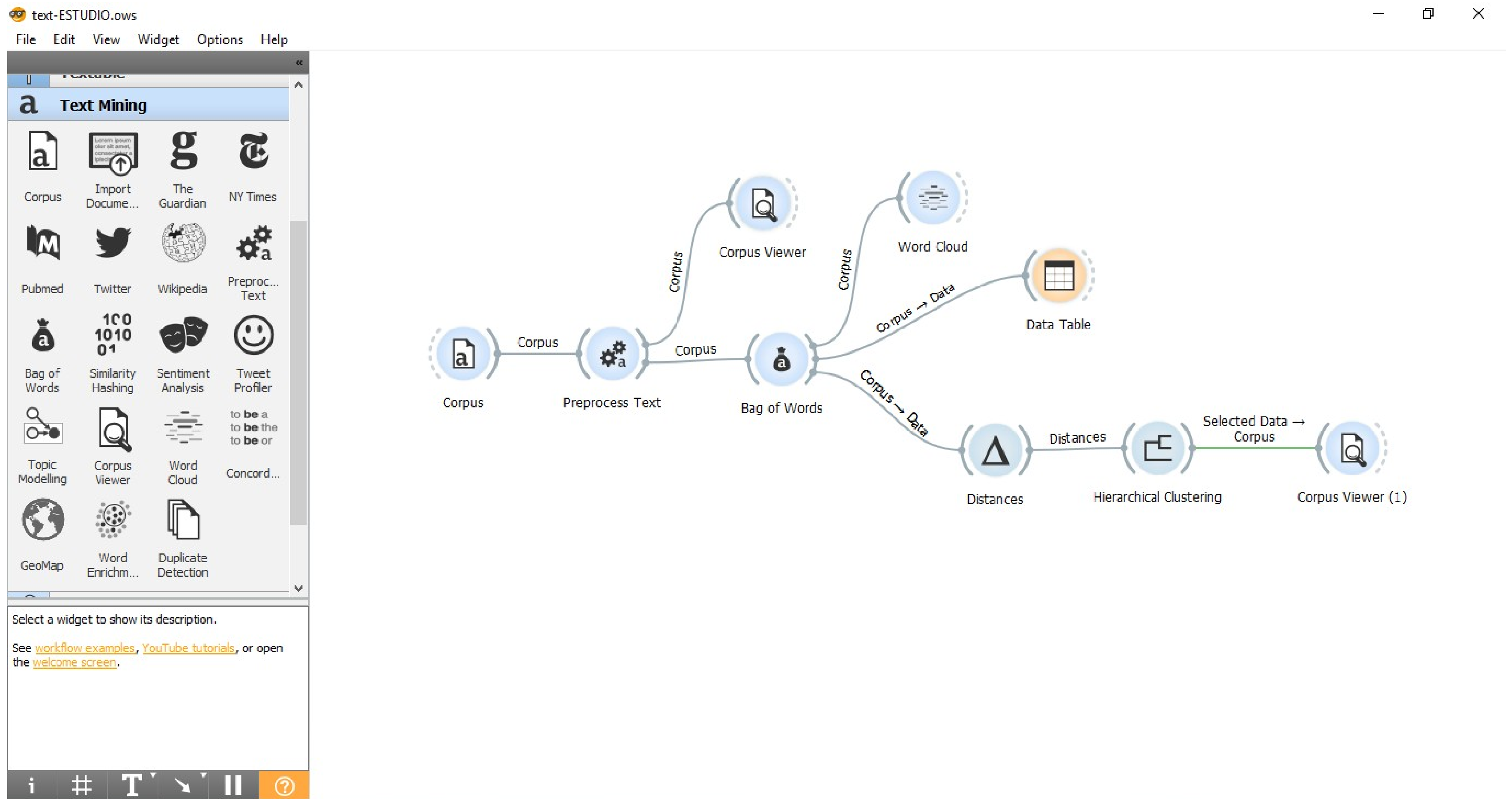
	EstadoPrestamo	Credito_ID	Dependientes	EstadoPrestamo(SVM)	Sexo
399	S	LP002978	0	S	Female
384	S	LP002917	0	S	Female
382	S	LP002894	0	S	Female
370	S	LP002813	1	S	Female
368	S	LP002804	0	S	Female
365	S	LP002794	0	S	Female
358	S	LP002757	0	S	Female
356	S	LP002753	1	S	Female
355	S	LP002741	1	S	Female
350	S	LP002731	0	S	Female

Trabajar con Texto
Working with Text

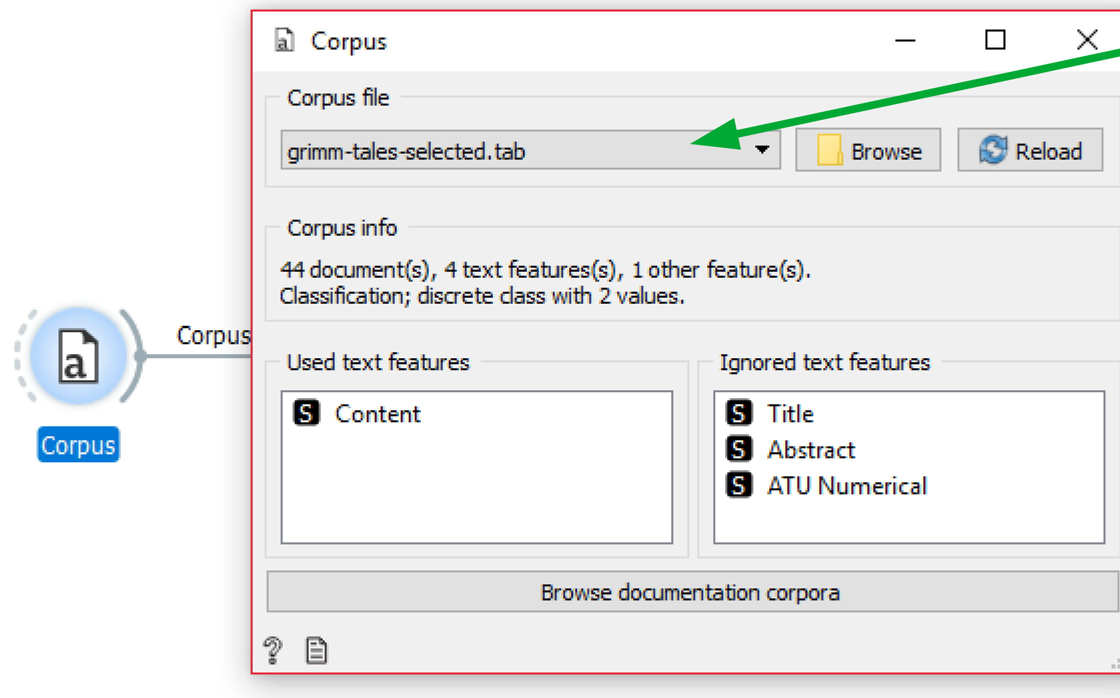
Material: **Software Orange,**

Fichero de datos: **grimm-tales-selected.tab** (es parte de software Orange).

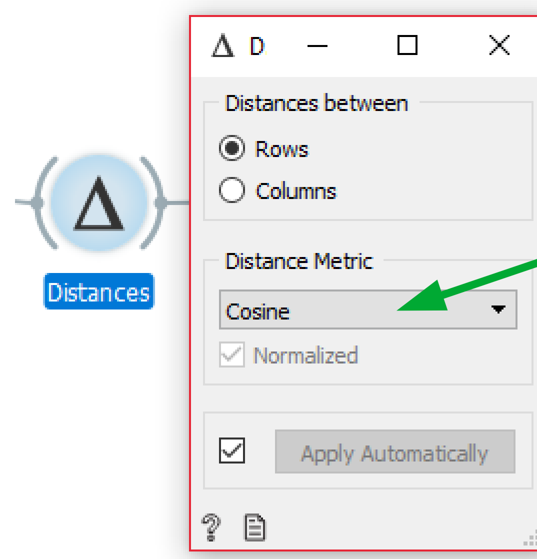
Creamos este espacio de trabajo:



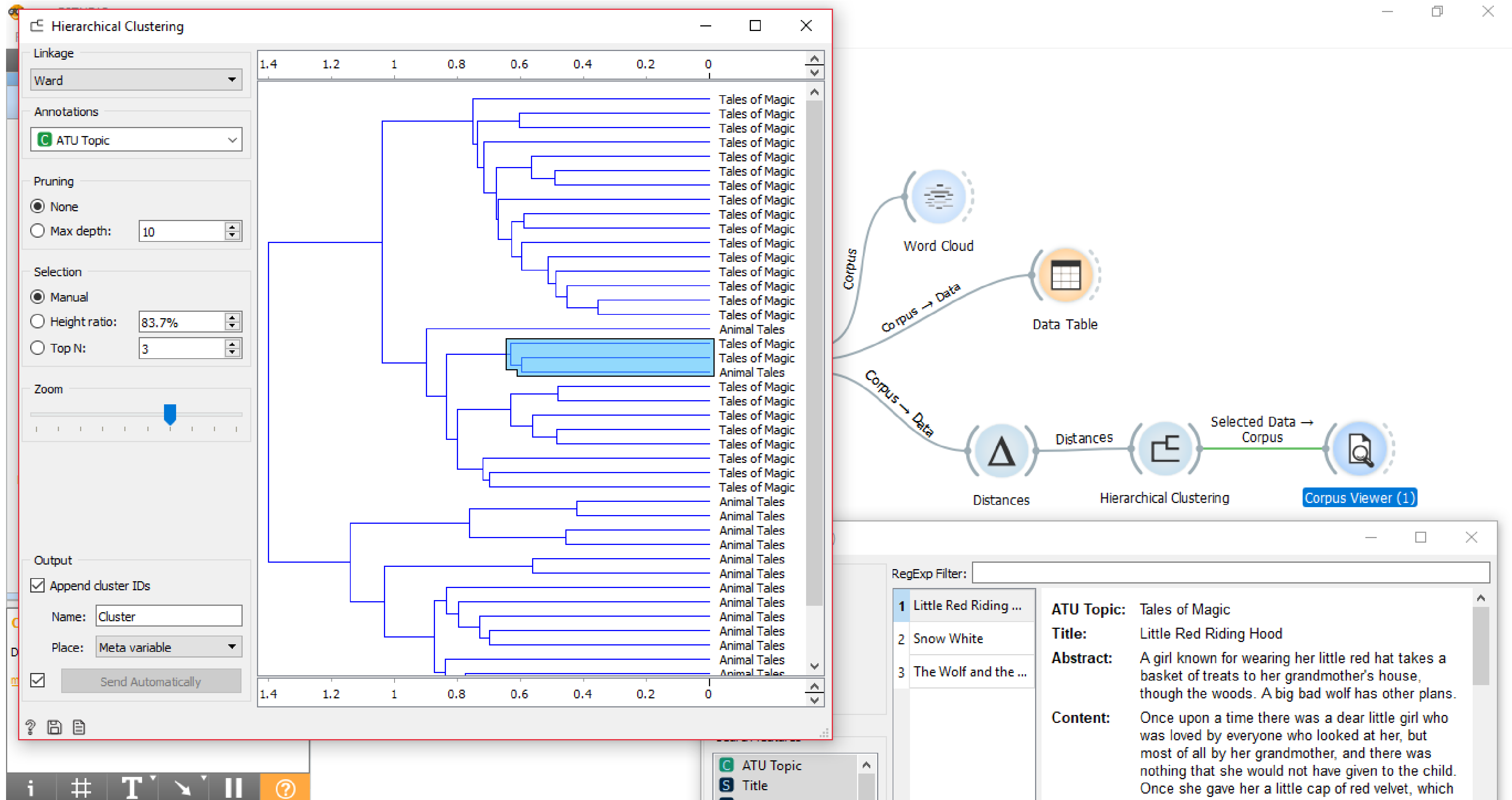
Y estos son los configuraciones:



En **Corpus** seleccionamos: **grimm-tales-selected.tab**



La métrica de distancia es en **Distances Cosine**



Ahora tenemos el análisis porque este dos documentos están declarados similares.

Este es solo un ejemplo de analisis de datos …

